Received: 25 November 2020 | Accepted: 17 March 2021

DOI: 10.1111/2041-210X.13608

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Automated retrieval of information on threatened species from online sources using machine learning

Ritwik Kulkarni[1] | Enrico Di Minin[1,2,3]

[1]Helsinki Lab of Interdisciplinary Conservation Science, Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland

[2]Helsinki Institute of Sustainability Science (HELSUS), University of Helsinki, Helsinki, Finland

[3]School of Life Sciences, University of KwaZulu-Natal, Durban, South Africa

Correspondence
Ritwik Kulkarni
Email: ritwik.kulkarni@helsinki.fi

Funding information
European Research Council, Horizon 2020, Grant/Award Number: 802933

Handling Editor: Phil Bouchet

Abstract

1. As resources for conservation are limited, gathering and analysing information from digital platforms can help investigate the global biodiversity crisis in a cost-efficient manner. Development and application of methods for automated content analysis of digital data sources are especially important in the context of investigating human–nature interactions.
2. In this study, we introduce novel application methods to automatically collect and analyse textual data on species of conservation concern from digital platforms. An end-to-end pipeline is constructed that begins from searching and downloading news articles about species listed in Appendix I of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) along with news articles from specific Twitter handles and proceeds with implementing natural language processing and machine learning methods to filter and retain only relevant articles. A crucial aspect here is the automatic annotation of training data, which can be challenging in many machine learning applications. A Named Entity Recognition model is then used to extract additional relevant information for each article.
3. The data collected over a 1-month period included 15,088 articles focusing on 585 species listed in Appendix I of CITES. The accuracy of the neural network to detect relevant articles was 95.91% while the Named Entity recognition model helped extract information on prices, location and quantities of traded animals and plants. A regularly updated database, which can be queried and analysed for various research purposes and to inform conservation decision making, is generated by the system.
4. The results demonstrate that natural language processing can be used successfully to extract information from digital text content. The proposed methods can be applied to multiple digital data platforms at the same time and used to investigate human–nature interactions in conservation science and practice.

KEYWORDS

biodiversity, digital conservation, machine learning, natural language processing, wildlife trade

1 | INTRODUCTION

Global biodiversity loss is one of the great sustainability challenges our society is facing (Butchart et al., 2010). Direct drivers of biodiversity loss are land and sea use change, climate change, pollution, invasive species and unsustainable harvesting (Maxwell et al., 2016). As gathering data on these direct drivers of extinction requires resources, which, in conservation science, are often inadequate, new technologies can be leveraged for this purpose (Pimm et al., 2015). In the Information Age, digital data can be leveraged to help address the global biodiversity crisis and study how humans interact with nature (Di Minin et al., 2015; Ladle et al., 2016). Methods for automated content analysis of this deluge of digital data are needed (Di Minin et al., 2019; Lamba et al., 2019; Toivonen et al., 2019).

Conservation culturomics is the field of conservation science where digital data sources and methods are being leveraged to help address the global biodiversity crisis and study human–nature interactions (Correia et al., 2021). Digital data sources have the potential to provide information on human–nature interactions at fine spatial and temporal scales (Di Minin et al., 2015). Digital data used in conservation science have thus far been mainly collected from webpages, social media platforms, video-sharing platforms, etc. (Ladle et al., 2016; Toivonen et al., 2019). There are multiple ways to access digital data and often more than one platform offers access to the same corpora. Data can be accessible through dedicated Application Programming Interface (API) services (e.g. Twitter API,[1]), online interfaces for data access (e.g. Google Trends,[2] GDELT[3]) and/or extracted using an automated script (Correia et al., 2021; Toivonen et al., 2019). An important aspect of digital data is that they allow for relatively low-cost research on often freely available data. However, attention should be paid to ensure responsible use of these data in accordance with data privacy requirements (Di Minin et al., 2021).

Applications of automated content analysis of digital data sources are increasing in conservation science (Toivonen et al., 2019). The application of algorithms for analysing visual, textual and/or audio content from digital sources can help study human–nature interactions at an unprecedented scale. These methods, for instance, are used for automatic identification, counting and description of species and individuals from images (Norouzzadeh et al., 2018). Frameworks and applications for the use of machine learning in conservation science are being promoted and used (Di Minin et al., 2018, 2019; Hernandez-Castro & Roberts, 2015; Lamba et al., 2019). However, the automatic extraction of information from text is still limited. Future developments should allow combining visual, textual and audio analysis of large volumes of data (Di Minin et al., 2019; Toivonen et al., 2019).

In this study, we introduce a novel application method for automated retrieval of textual information pertaining to species of conservation concern from digital platforms. Specifically, we focus on retrieving information on 791 species listed in Appendix I of CITES from online news sources. Appendix I includes the most endangered among CITES-listed species that are threatened with extinction where international trade in specimens of these species is prohibited except when the purpose of the import is not commercial. We focus on online news because they contain valuable data, such as information about which species are traded and confiscated by authorities, that can be potentially used to monitor the trade in Appendix I listed species. To our best knowledge, fully automated methods for extracting relevant information for a wide range of species in a single collection effort are still missing in conservation science. Without adequate filtering techniques allowing for the identification of relevant information only, it would be unfeasible to manually classify content on hundreds of species from multiple digital platforms globally (see Stonebraker et al., 2013, for a general discussion on automated curation). Previous studies either manually collected information on certain species (see e.g. Bombieri et al., 2018; Corbett, 2016; Morcatty et al., 2020; Nghiem et al., 2016; Unger & Hickman, 2020) or manually reviewed reptile selling websites after using search engines to identify these websites (Marshall et al., 2020). Here, we introduce an automated system that not only collects online news articles related to species of conservation concern but also uses machine learning tools and natural language processing to filter relevant content. Importantly, we use text vectorisation methods to deduplicate articles and to train a neural network that learns to classify relevant articles from irrelevant ones. For each article, we extract named entities using natural language processing and enhance the information retrieval potential of the database.

2 | MATERIALS AND METHODS: PIPELINE

We describe in detail the stages of a complete end-to-end pipeline that initiates a web search for online news articles related to CITES Appendix I listed species, and proceeds to filtering, classifying and extracting information from the downloaded articles. The general overview of the pipeline with its main stages is depicted in Figure 1. Here, we list the main flow of the pipeline and highlight some of the important features while the rest of the section elaborates on each of them and we refer the reader to the subsections for specific details:

1. *News collection*: News articles were compiled together from two channels, namely, Google News and Twitter. Google News was targeted towards the species listed in CITES Appendix I while specific Twitter handle postings about wildlife trade were targeted for selecting conservation-related topics (not limited to specific species). All downloaded articles were collected within a MongDB database and amended with additional information related to the species in question (Refer Sections 2.1.1 and 2.1.2).

[1] https://developer.twitter.com/en/docs/twitter-api/

[2] https://trends.google.com/trends/

[3] https://www.gdeltproject.org/

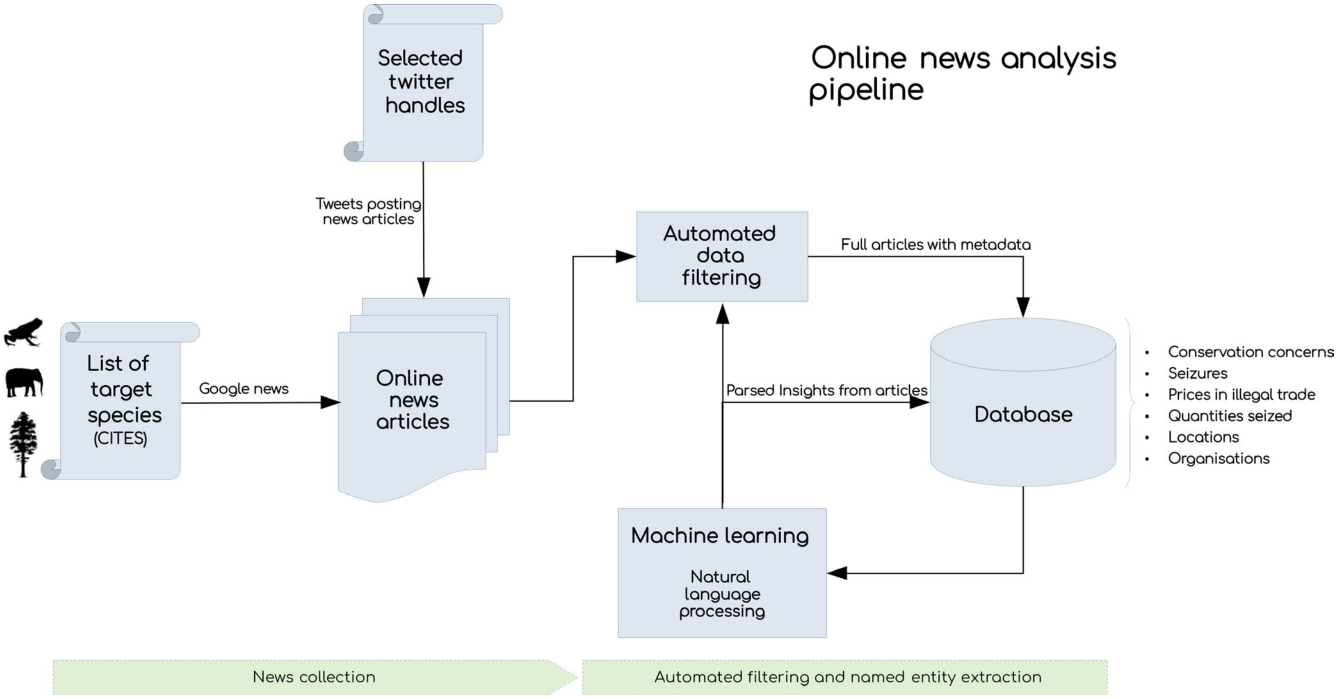

Online news analysis pipeline
Selected twitter handles
Tweets posting news articles
List of target species (CITES)
Google news
Online news articles
Automated data filtering
Full articles with metadata
Parsed Insights from articles
Database
Conservation concerns
Seizures
Prices in illegal trade
Quantities seized
Locations
Organisations
Machine learning
Natural language processing
News collection
Automated filtering and named entity extraction

FIGURE 1 Framework for collecting and automatically filtering information on species listed in Appendix I of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) from online news articles and Twitter

2. *Automated data filtering*: After the collation of articles, all data were filtered for duplicates and irrelevant articles using machine learning techniques. Deduplication was performed using a vector comparison algorithm and thus takes into account text similarity of articles rather than just webpage links or titles. Meanwhile, irrelevant articles were removed using a neural network trained to identify such articles from relevant ones. A crucial aspect at this stage was the method for automatic annotation of training data, which can be a stumbling block in many machine learning applications (Refer Sections 2.2.1 and 2.2.2).
3. *Named entity extraction*: Once the articles are selected, we apply Named Entity Recognition to tag articles and particular sentences with specific information such as names of persons, prices, locations, etc. Thus, for each article in the database, there exists parsed information that can be accessed for further fine-grained analysis that may assist with research (Refer Section 2.3).

The pipeline (scripted in Python 3.6, Van Rossum & Drake, 2009) is automatically triggered every month to have a regular update of the data. Throughout the execution of the pipeline, we implement parallel processing of various stages. Parallel processing allows us to execute different computational tasks simultaneously on separate Central Processing Units (CPUs), rather than sequentially processing each task, thus significantly reducing the total time of pipeline execution. Since the processing of each news article is independent (different articles have no dependence on each other to be analysed at various stages of the pipeline), the processing then falls under the category of '*Data Parallelism*' (Régin et al., 2013) where each process can be split into identical copies of itself working on different subsets of data in parallel. This leads to a significant gain in efficiency and reduces total execution time ~4 times when using five threads of the CPU. Each complete execution of the pipeline requires about ~560 min under the configuration as described in the following sections. Parallel processing is implemented using the Python package Joblib.[4]

2.1 | News collection

As a proof of concept, online news articles were collected using two channels, Google News (https://news.google.com/) and Twitter (see Figure 2). Google News was selected as it is a popular news aggregate site that provides a search functionality based on both keyword and query while listing a continuously updated news database from around the world. Similarly, another popular source of information is social media and we target specific wildlife-trade-related handles on Twitter (see Sections 2.1.2). The advantage of using Twitter handles was the high likelihood of topic-relevant articles since they are manually posted by moderators of the handle.

2.1.1 | Google News search

Search terms for the Google News search engine were composed using the list of species from CITES Appendix I that can be

[4] https://joblib.readthedocs.io/en/latest/

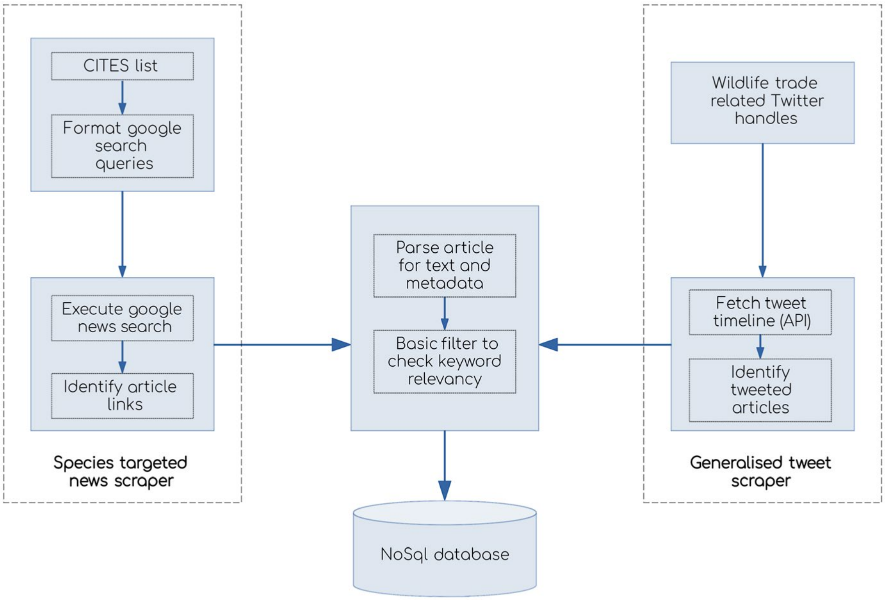

CITES list
Format google search queries
Execute google news search
Identify article links
Species targeted news scraper
Parse article for text and metadata
Basic filter to check keyword relevancy
NoSql database
Wildlife trade related Twitter handles
Fetch tweet timeline (API)
Identify tweeted articles
Generalised tweet scraper

FIGURE 2 Flow diagram of the stage to collect articles on species listed in Appendix I of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). Information from two channels (Google news and Twitter) is collated into a single structured database

downloaded from the CITES website.[5] We searched by the English common names of the species (listed in the CITES data), as we were searching for news articles in the public domain that are primarily intended to be read by the public, but index them by the scientific name in our database. The Python package BeautifulSoup (Richardson, 2007) was used to parse and obtain the HTML source of the search result for each of the species. Hyperlinks to the news articles were then identified within the HTML source with '*./articles*' tag in Google News URL (Uniform Resource Locator) and collected as a potential source of news articles. Google News limits the number of search results to 100 latest articles. Refer to Figure 2 for the flow diagram.

Once all links to the articles were obtained, the main text and metadata for the links were parsed using the package Newspaper3k.[6] The fields of interest are shown in Figure 3 under 'Article Metadata'. We iterate over all the species in the list using their stated common names and collate all the found articles. The data were then supplemented with species-related information so that the final database entry for each of the species contains the fields depicted in Figure 3 as 'Document Fields'. The International Union for Conservation of Nature (IUCN[7]) threat category information was appended from a separate list that connects the scientific name to the corresponding IUCN Red List category and when not available, listed as 'NA'. Each entry was then written in a NoSql format to a local MongoDB data server (Banker, 2011).

2.1.2 | Twitter handles

Twitter is used by many organisations and institutions to publicly share relevant information regarding several issues. We selected three Twitter handles related to wildlife trade, which are known to tweet information about wildlife trade. The three handles were '@TRAFFIC_WLTrade', '@IlWildTrade' and '@WLTradeNews' (this list can be easily extended to include other handles). Figure 2 depicts the flow diagram for collecting news articles from Twitter and is linked to the process described in Section 2.1.1.

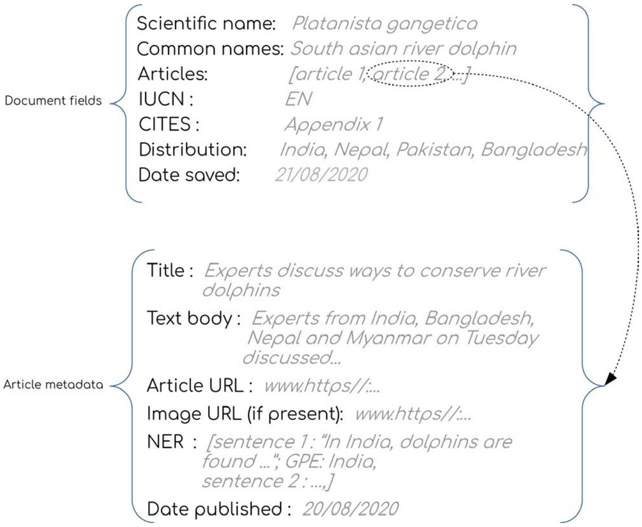

Document fields
Scientific name: Platanista gangetica
Common names: South asian river dolphin
Articles: [article 1, article 2, ...]
IUCN : EN
CITES : Appendix 1
Distribution: India, Nepal, Pakistan, Bangladesh
Date saved: 21/08/2020
Article metadata
Title : Experts discuss ways to conserve river dolphins
Text body : Experts from India, Bangladesh, Nepal and Myanmar on Tuesday discussed...
Article URL : www.https//:...
Image URL (if present): www.https//:...
NER : [sentence 1 : "In India, dolphins are found ..."; GPE: India, sentence 2 : ....]
Date published : 20/08/2020

FIGURE 3 Information collected at the document level (Document Fields) and the metadata for each article extracted (Article Metadata) for species listed in Appendix I of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)

We access the Twitter API[8] using the Python package Tweepy (Roesslein, 2020) and use the Cursor Object in Tweepy to obtain the Twitter timeline of the desired handle. A Cursor Object helps to loop through sequential pagination (discreet set of digital pages that host

[5] http://checklist.cites.org/#/en

[6] https://github.com/codelucas/newspaper/

[7] https://www.iucn.org/resources/conservation-tools/iucn-red-list-threatened-species

[8] https://developer.twitter.com/en/docs

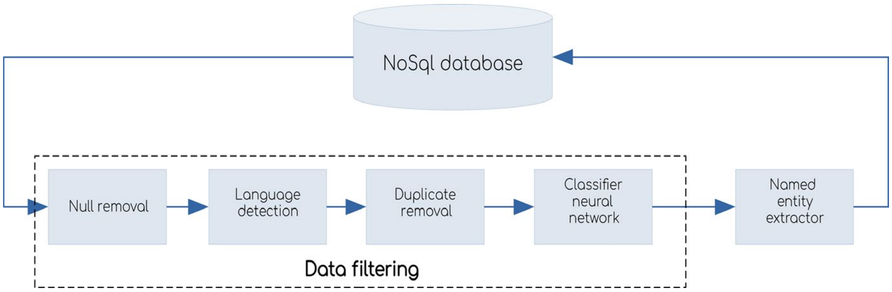


FIGURE 4 Post-data download processing stages depicting the flow of the four filtering stages and the stage of extracting named entities

the tweets) on Twitter and make requests for tweets on each loop. The Twitter API limits the number of tweets to 200 per request. However, by making successive requests for older tweets, we were able to download a total of 3,200 tweets (maximum allowed) per handle in a single run. After obtaining the tweets, we identify whether any hyperlinks to external articles were present in the tweets and collect all such links. Following the link collection, the remaining steps are the same as described in Section 2.1.1 that lead to the database entry.

2.2 | Automated data filtering

Data collected from search engines like Google News result in a mixture of relevant and irrelevant content, requiring further filtering. For example, articles that contain a targeted keyword, but, for example, contain information on businesses named after a species that is also listed in CITES Appendix I, are irrelevant and should be discarded. The process of scanning and removing articles was done offline as it can be resource heavy and time-consuming. Data filtering was done in four main stages (Figure 4), namely, (a) null article removal, (b) language identification, (c) duplicate removal and (d) articles' selection after classification using a neural network. At the end of each run, we maintain a log of how many articles have been filtered out at various stages.

The first stage of filtering was simply the removal of null entries in the database, either the entry of a species that has no articles listed for it or an article entry which has no text body. Such entries usually arise from spurious results of fetching the HTML data due to website-specific reasons or server restrictions. The second stage involved removal of articles that are not in English. Non English articles, although rare, tend to arise from Twitter handles. Language of the article was detected using the Python package LANGDETECT[9] and all articles not in English were removed from the list. Removal of non-English articles was needed due to downstream algorithms, namely, the classifier and named entity recogniser. These models are trained on English data and currently work only on English text. Single models handling multiple languages are currently being developed and have yet to achieve strong performance on a wide array of languages (Toleu et al., 2020). Therefore, models for each required language are still needed. The remaining two stages require more elaborate processing and are described in the following sections.

2.2.1 | Duplicate removal

On a number of occasions news articles can have duplicates. We ensure that a particular URL was never repeated for extracting an article. However, due to the nature of the media, many news sources publish the same article with little or no modification (termed as 'syndicated' copy). We deduplicate such articles using a vector comparing operation. Deduplication was performed on per species level for each entry in the database, meaning that an entire batch of articles listed under a particular species was considered for deduplication at a time and all articles for each species was successively deduplicated. Hence, an article will not repeat content for a species, but may occur again as an entry for another species, as a single article can cover several species. This helps us maintain species-level statistics.

To perform deduplication, we first vectorise the text using Tf-Idf (Salton, 1991), as implemented in the Python package SCIKIT-LEARN (Pedregosa et al., 2011). Tf-Idf is a simple yet effective method to convert text into numerical vectors, achieved by term weighting words based on their frequency of occurrence in a document and then normalising the weight by the popularity of the word across the entire corpus (in this case, all articles for that particular species). Vectorising allows us to perform mathematical operations on the text, which we used to evaluate duplicates. Each article was indexed by a unique reference id. The text of the entire article was vectorised with all the content without any pre-processing to remove words. This helps represent the full information in the Tf-Idf vector. Vectorisation is repeated for the whole set of articles after which we calculate the cosine similarity of the article vectors as

$$\cos(\theta)_{ij} = \frac{A_i \cdot A_j}{|A_i|\,|A_j|}; \quad i, j = 1, 2, \ldots, n. \tag{1}$$

A_i's are the article vectors indexed from 1 to n. The cosine value spans from 0 to 1, where 1 indicates identical vectors, whereas 0 would mean no match between vectors. Equation 1 gives a $C_{n \times n}$ matrix of cosine pairs and after removing the diagonal (pairs with self), we collected all pairs with cosine ≥0.95. These pairs indicate that the vectors were very similar and hence the articles were near identical. Since

[9] https://pypi.org/project/langdetect/

the matrix is symmetric, we only consider one half of the matrix to speed up the operation. Upon manual inspection of 50 articles and their near identical articles, it was seen the slight reduction from 1 in identical articles was due to minor modification of a few lines in syndicated feeds. The threshold for duplicates was selected heuristically by observation, after determining the cosine range that is spanned by pairs of duplicates and confirming that those below the threshold are not identical. The articles with high cosine but below the threshold were articles that discussed the same event but not a syndicated copy. After knowing the duplicate ids of the articles, only one article was retained while the rest were removed from the list. It is important to note that this method removes duplicates where a large body of text is identical. However, some original articles discussing the same event, but naturally differing in nature of expression (using a different vocabulary and linguistic style) may still be retained.

2.2.2 | Neural network classifier

As mentioned above, news articles retrieved from Google News were not always relevant. Given that frequent manual curation of large sections of the database is not only time-consuming but also resource-intensive, we developed an automatic way of pruning out irrelevant articles using a neural network. Neural networks have become quite powerful at classifying text (Kowsari et al., 2019; Kulkarni et al., 2018) in the past decade. The following sections describe the procedure of building, training and testing a neural network to perform a classification task on the collected articles.

Architecture

A four-layer feed-forward network makes a binary prediction whether the input text was relevant or not (in the context of biodiversity conservation; Figure 5). The neural network was built using Keras[10] with a Tensorflow backend (Abadi et al., 2016; these are Python modules to build mathematical models and train neural nets). The input to the model was in the form of a 1,000 dimensional vector obtained from transforming the text using a Murmurhash3 hashing function (Dahlgaard et al., 2017) implemented in Scikit-learn. A hashing function is a method to convert text into vectorised arrays and has several advantages like not requiring to maintain a specific vocabulary along with needing low memory usage. This helps scale the model as new articles are added on each cycle of data collection and the vocabulary grows. With a sufficiently large size of the Hashing vector, the collisions, a term denoting the incidence of different words being mapped to same vector rather than different ones, can be significantly reduced. The size of the hashing dimension was selected as the minimum size that does not hamper learning performance of the model. The size of the hidden layers in the neural network was successively reduced with dimensions as depicted in Figure 5. All layers except the final prediction layer used rectified linear activation function ($f(x) = x^{+}$). The prediction layer, which was a single unit layer, used a sigmoid activation function ($f(x) = \frac{1}{1+e^{-x}}$), making the prediction bounded by 0 and 1 (see Supporting Information for the model summary). We used the Dropout method, in which a random fraction of units are removed during training to regularise the model (Srivastava et al., 2014) and reduce overfitting. The network was trained using a binary cross-entropy loss function (Ramos et al., 2018) for 10 epochs with a batch size of 50 and the iteration with the best performance on the validation set is saved as the final model. The metrics used to evaluate model performance were Accuracy, Precision, Recall and *F*-score (see Goutte & Gaussier, 2005 for a discussion on the metrics).

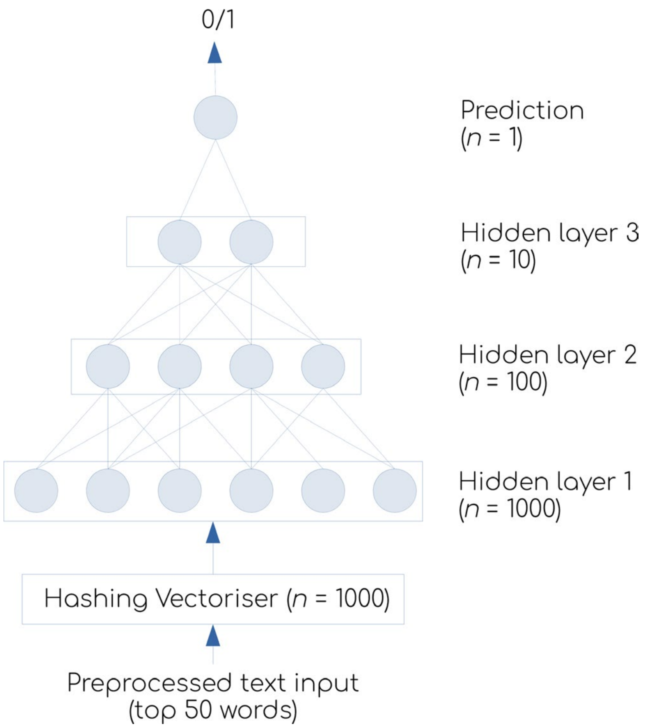


FIGURE 5 Neural network architecture depicting a binary classifier for articles to determine relevancy

Training

The model needs to be trained to understand the difference between relevant and irrelevant articles. As such, it requires training examples to distinguish between the two categories. Thus, as a supervised learning algorithm, it requires labelled training data for the model to adapt its parameters. In many cases, labelled data for the specific purposes of a study do not exist, and/or can be difficult to obtain, and/or require investing time and resources in building an annotation setup. To address these limitations, we developed a novel approach to implement an automatic labelling scheme for the data. To train the classifier, we need two sets of articles: (a) relevant articles (concerning conservation issues) and (b) irrelevant articles (general news). Articles relating to biodiversity conservation were sourced from Twitter handles and, as these are manually posted, they predominantly pertain to conservation topics. All articles sourced from Twitter were assigned the label of 'Relevant'. General news articles (not specifically discussing conservation issues) were sourced from two general news

[10] https://keras.io/

TABLE 1 Performance metrics of the neural network model on test data

Accuracy	Precision	Recall	*F*-score
95.91	93.75	98.95	96.28

datasets: (a) The Reuters corpus of news available through NLTK (a natural language toolkit in Python; Bird et al., 2009) that contains 10,788 articles spanning 90 different topics and (b) '*all the news*' dataset hosted on the Data Science resource website Kaggle.[11] This dataset contains over 100,000 articles covering various news sources. Articles were sourced in equal proportion from both datasets using random sampling. We used two different general news datasets to cover a wider range of topics, thus giving a better representation of general news articles. Conservation-related articles numbered 2,464 while the 'irrelevant' articles were selected to be 3,000 so as to balance the proportion binary labels assigned to each group for training. The labelled data were split into 80% for training and 20% for testing purposes. The input to the model for each article consisted of preprocessed text which entailed removing of stopwords (high-frequency words that do not possess any content and hamper the performance of the model), removing special characters and numbers, then finally selecting the top 50 words from each article based on their frequency of occurrence in the article. All words were converted to lower cased letters. The list of stopwords was obtained from NLTK with a few custom additions such as: '*told*', '*Said*', '*According*', etc. (see Supporting Information for a complete list). These words were found to be high frequency and common for news articles. The model showed a high classification performance on the test set as shown in Table 1. This indicates there was enough signal to be differentiated between the texts of the two classes and at the same time gives credence to the automatic data labelling process.

Evaluation

We look at the performance of the model on the articles extracted using the CITES List and Google News. These articles are neither part of the training data nor from the same source. Hence, it is important to evaluate the model performance on this data, as it would constitute the target domain for relevancy filtering. We passed 50,256 articles collected from the Google News search through the trained model and analysed the model prediction. The model prediction lies between 0 and 1 with 0.5 as the prediction decision boundary (≥0.5 set to 1; 0 otherwise), 0.5 is also the point of maximum ambiguity for classification. Hence, to reduce ambiguity for relevant predictions, we set 0.4 as the threshold below which articles were considered relevant and above 0.4 as irrelevant. This had the effect of having a high likelihood for true positives at the cost of losing some false negatives. In the interest of cleaning the data, we believe this to be an optimal choice because it retains less number of articles erroneously classified as relevant. For an aggregate analysis, we measure the top frequency words in either of the predicted classes shown in Table 2 (top 15 words). A total of 40,863 (81.4%) articles were classified as relevant while 9,393(18.6%) as irrelevant.

TABLE 2 Aggregate frequency analysis of model predicted classes on extracted articles (top 15 words)

Predicted relevant	*species, wildlife, conservation, endangered, animals, wild, world, found, national, animal, like, population, park, forest, people*
Predicted irrelevant	*like, people, old, world, around, back, species, see, park, zoo, national, animals, us, long, get*

Table 2 shows that the model broadly captures the difference between conservation and irrelevant news articles, as can be seen from the type of high-frequency words across all articles within the class. However, the presence of words such as '*species*' and '*animals*' in the irrelevant section suggest relevant articles have a tendency to get misclassified more often than irrelevant articles being misclassified as relevant. Even so, the occurrence of '*species*' was ~2.29 times less in proportion in the irrelevant articles class compared to relevant articles class while '*animals*' was ~1.36 times less. List of absolute frequencies are shown in Supporting Information. Upon closer inspection, it was seen that since the model is trained with conservation focus articles, certain type of animal, or plant-species-related articles, which are about zoos, museums, local animal theme events or wildlife photography, did not get classified as irrelevant, while only a few articles that we consider should have been classified as relevant got marked as irrelevant. As we gather more data, we aim to retrain the model with a larger dataset to improve the real-world performance of the model.

2.3 | Named entity extraction

We supplement the database with an additional feature of named entities to help with analyses of the articles. This can be useful to get specific information as described in Section 3.3, for example, to get reported prices and quantities of traded animals. Named entity recognition (NER) is the process of locating and classifying words or phrases in a plain text sentence, into categories such as '*Persons*', '*Places*', '*Events*' and so on. The text of each article was scanned for named entities and corresponding sentences were extracted out of the text. We used the Python package spaCy,[12] which provides several models and methods to perform various natural language processing tasks, with NER being one of them. For the purpose of this study, six types of entities are extracted (Table 3) using the *en_core_sm* model for English. This model has an *F*-score[13] of 85.55 on the NER task it is trained for. NER models are likely to lose their performance when applied to a different domain but even with

[11] https://www.kaggle.com/snapcrack/all-the-news

[12] https://spacy.io/

[13] https://spacy.io/usage/facts-figures

identification errors the results still remain largely useful for analysis purposes. Thus, for every article in the database, there was corresponding information for that article about the named entities (type and text) in the article along with the particular sentence in the article that contains those respective entities. This was the last stage of data processing and the document entry for each species was finalised after the data filtering and NER extraction stages.

TABLE 3 Tags extracted for Named Entity Recognition, as defined in spaCy and their corresponding descriptions

spaCy entity extracted	Description
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
GPE	Countries, cities, states
PERSON	People (names)
ORG	Companies, agencies, institutions, etc.
CARDINAL	Numerals

All the collected data were stored in a NoSql database format (Han et al., 2011) using MongoDB. MongoDB is a distributed database that stores data in a flexible format (no strict scheme for data storage and uses JSON like documents, Pezoa et al., 2016).

3 | RESULTS

3.1 | Data description

The pipeline resulted in the collection of 13,565 news articles mentioning a total of 585 species in CITES Appendix I. In addition, 1,523 articles were collected from the three Twitter handles (@WLTradeNews: 797, @TRAFFIC_WLTrade: 486 and @IIWildTrade: 240), taking the total to 15,088 articles. From a machine learning point of view, the corpus size of the dataset is 11,134,822, while the number of unique tokens is 419,706. As is the case in many real-world data distributions, the frequency of the articles per species shows a distribution similar to a power law with an exponential like decay (Figure 6: inset). Top 30 species from the total distribution are shown in the main plot with lion *Panthera leo* having the most

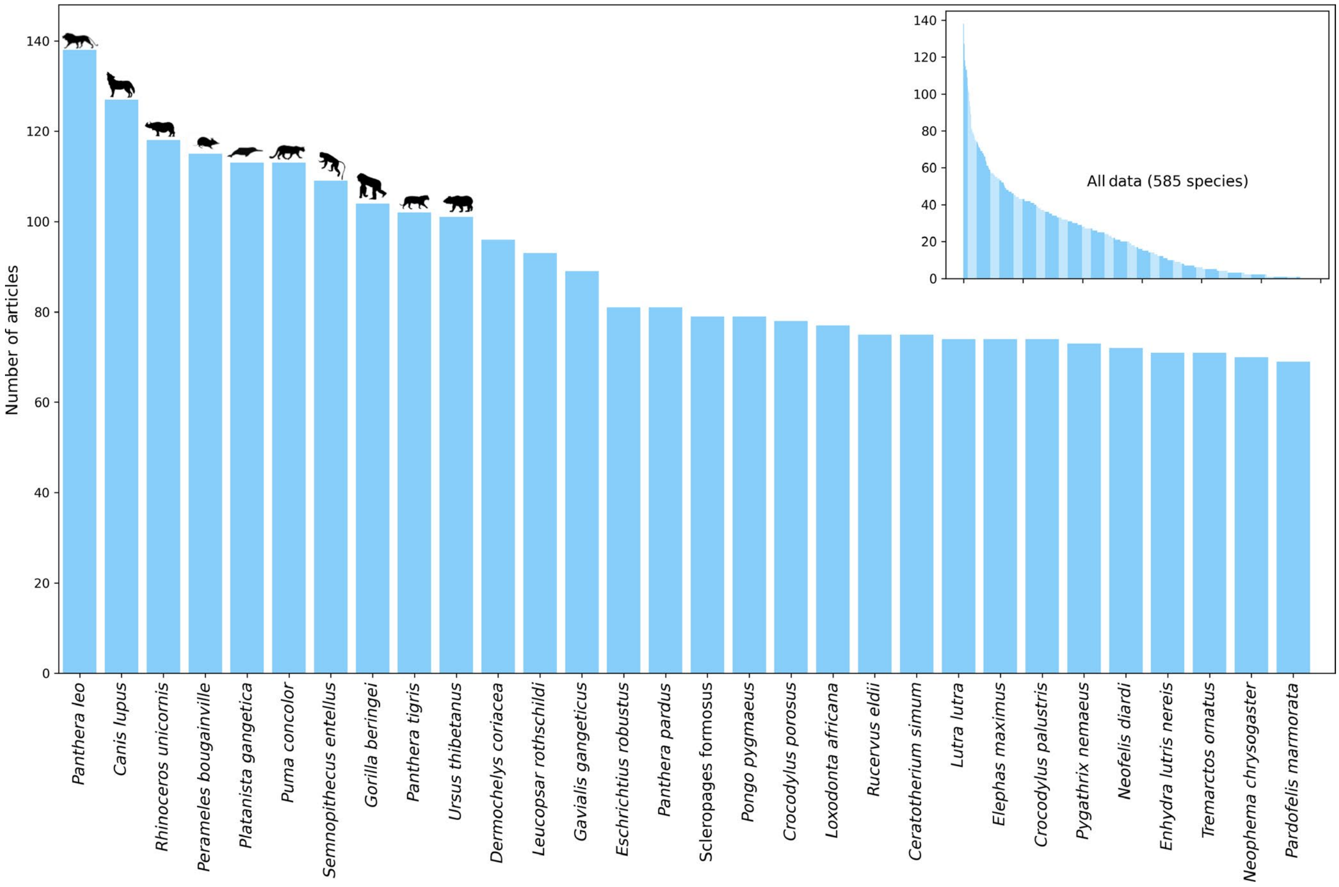


FIGURE 6 Number of articles found for species (top 30 species in main plot and all species in inset) listed in Appendix I of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)

number of articles. Charismatic species such as wolf *Canis lupus*, greater one-horned rhino *Rhinoceros unicornis*, elephant *Elephas maximus* and *Loxodonta africana*, orangutan *Pongo pygmaeus*, grey whale *Eschrichtius robustus*, etc., species of conservation concern such as gharial *Gavialis gangeticus*, Bali myna *Leucopsar rothschildi*, western barred bandicoot *Perameles bougainville* or common species like grey langoor *Semnopithecus entellus* were the most covered in the database. It is important to note that the frequency count depends on the mention of the species in the article and may not always mean that the whole article discusses only the one species in question.

Figure 7 shows the proportions of taxonomic groups found in the database. Animals are covered more than plants and mammals and birds are covered more than other groups in the database. The focus on mammals can be partly explained by the over-representation of this class in Appendix I of CITES (see Supplementary Materials: Figure S2). However, some other classes appear more covered or less covered in online news regardless of the number of species included in that class. To explore this, we normalise the articles proportion by the number of species for that class in the database. Normalised numbers in Table 4 show

TABLE 4 Ratio of the number of articles per class normalised by the proportion of species in that class. A ratio of 1 for a class signifies that the number of articles for that class simply reflects the proportion of species present in the class, that is, the number of articles for each species in the class is equal to the mean number of articles extracted per species across the database, while any variation around 1 would either mean an above average number of articles for that class based on the proportion of species (if >1) or below average (if <1)

Class	Articles ratio
Mammalia	1.22
Aves	0.82
Reptilia	0.90
Actinopteri	1.75
Amphibia	1.0
Elasmobranchii	1.0
Coelacanthi	2.0
Bivalvia	0.13
Insecta	0.4

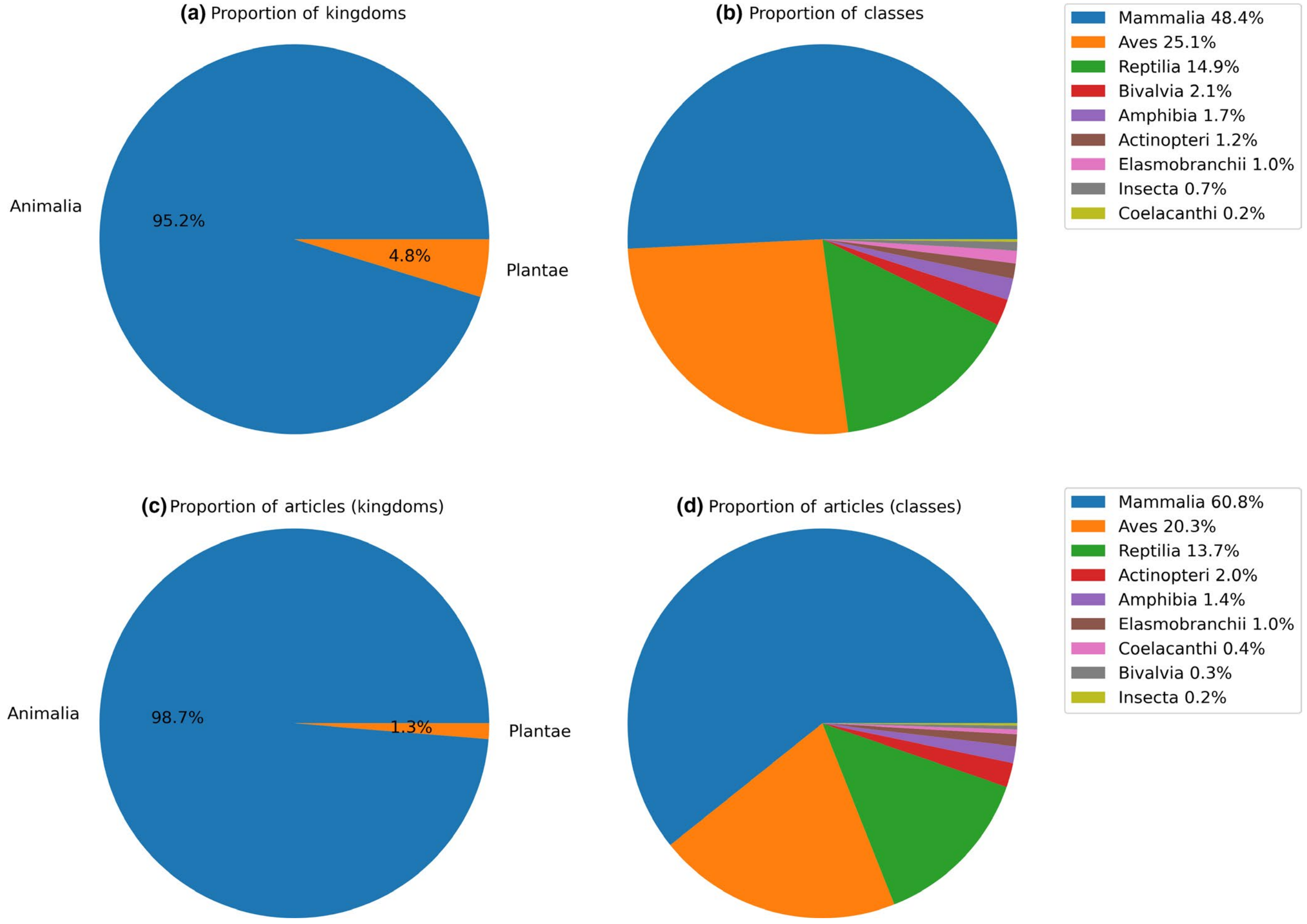


FIGURE 7 Proportion of taxonomic groups and articles covering these taxonomic groups. Top row: proportion of species by (a) kingdom and for (b) animal classes present in the database. Bottom row: proportion of articles by (c) kingdom and for (d) animal classes

that the Coelacanthi class has twice the mean number of articles across the database while Bivalvia has the lowest with slightly over 1/10th the mean.

As a gross indication of the content in the news articles for some of the species, we show wordclouds (Figure 8) for four species *Gorilla beringei* (eastern gorilla), *Dermochelys coriacea* (leatherback turtle), *Pezoporus wallicus* (eastern ground parrot) and *Dalbergia nigra* (Bahia rosewood) that span four different taxonomic groups. Words such as '*critically endangered*', '*conservation*', '*habitat*', '*fire*' (re: Australian bushfire), etc., indicate the content of the news articles.

3.2 | Brief overview of articles extracted from Twitter

Articles extracted from Twitter originate from various tweets on the timeline of the Twitter handle. Hence, they are not already tagged or categorised into specific species or topics. Therefore, we use topic modelling to discover the topics mentioned in the articles (see Supporting Information for details about the topics model). The words composing each of the topics are listed in Table 5 along with the abstract label as interpreted by the authors. Table 5 gives a sense of what the Twitter extracted articles cover in content.

3.3 | NER utility examples

We give a few brief examples of how the NER feature can help extract useful information related to the species of interest. As, for all articles there are individual sentences marked with the named entities, it allows us to query the database regarding specific entities and species. For example, if we want to explore information about prices involved in the trade of pangolins, we can query the database using the spaCy entity identifier for 'MONEY' listed in Table 3 along with the keyword '*pangolin*'. An example of Named Entity extraction is shown in Figure 9. At present level, overall named entity figures are as follows: first number is the total number of sentences extracted while the second number in brackets is the number of

TABLE 5 Topics indicating the content of Twitter extracted articles

Topic number	Defining words	Abstract label
1	*animal*, *species*, *wild*, *tiger*, *trade*, *pangolin*, *wildlife*, *bird*, *population*, *find*	Asia focused wildlife trade
2	*conservation*, *people*, *work*, *community*, *area*, *make*, *project*, *human*, *world*	Community-based conservation
3	*illegal*, *government*, *report*, *forest*, *fish*, *environmental*, *log*, *abalone*, *country*, *law*	Legal issues
4	*wildlife*, *trade*, *ivory*, *illegal*, *elephant*, *country*, *international*, *traffic*, *china*, *product*	Africa focused wildlife trade
5	*wildlife*, *police*, *arrest*, *seize*, *poacher*, *case*, *official*, *rhino*, *man*, *department*	Law enforcement focus wildlife trade

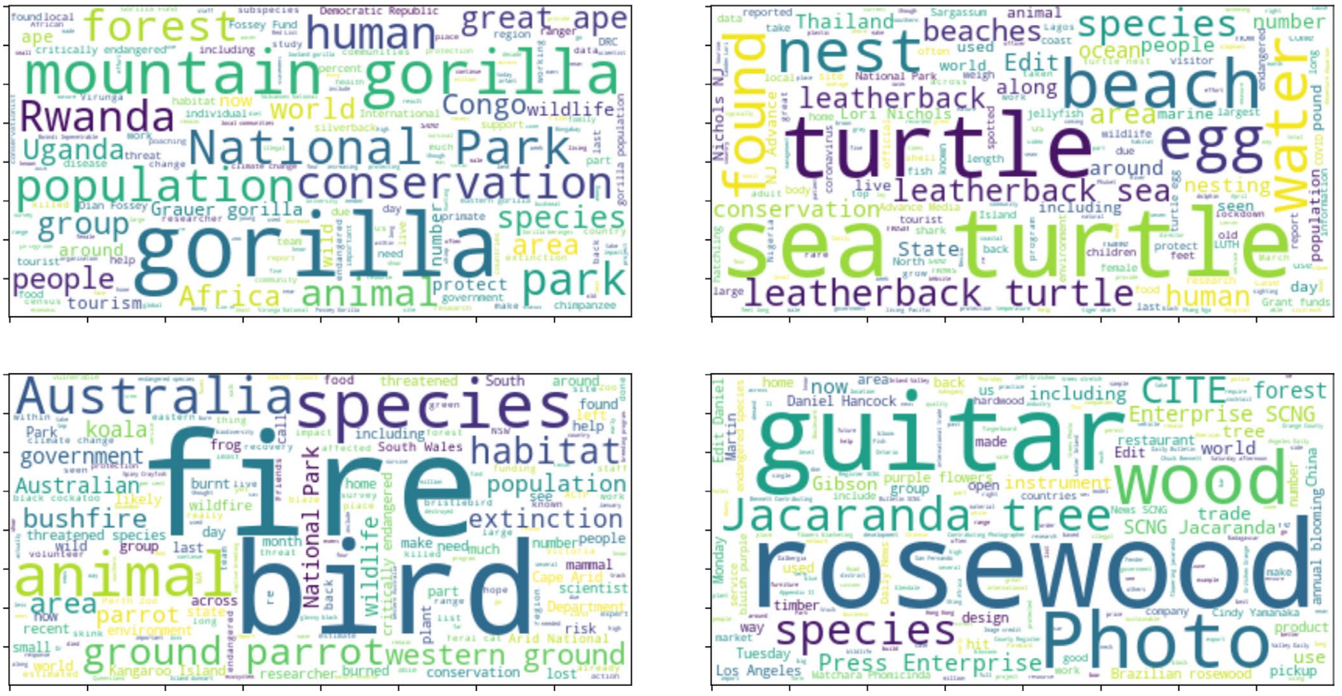
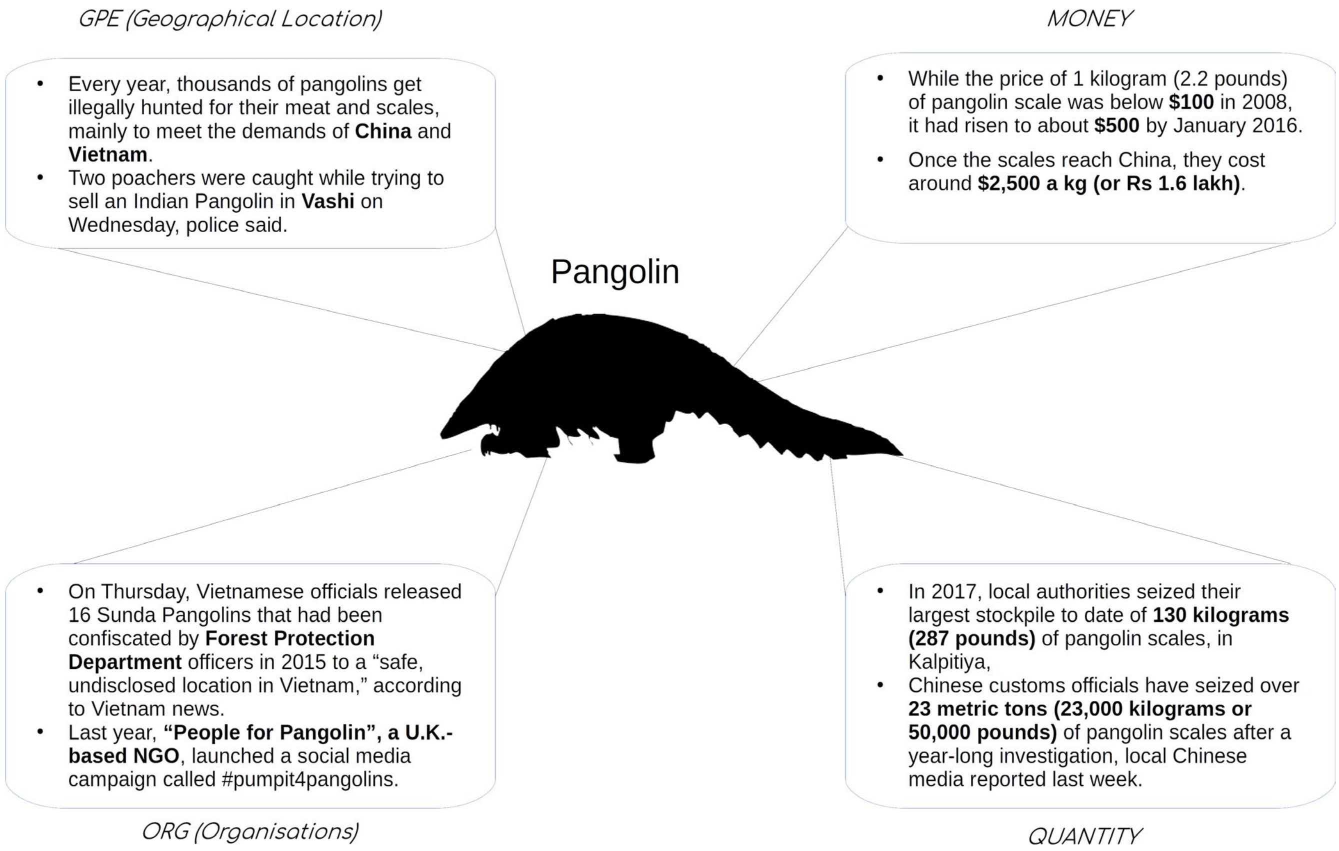


FIGURE 8 Wordclouds of all news articles for four species, namely, eastern gorilla *Gorilla beringei*, leatherback turtle *Dermochelys coriacea*, eastern ground parrot *Pezoporus wallicus*, Bahia rosewood *Dalbergia nigra*, listed in Appendix I of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)



- Every year, thousands of pangolins get illegally hunted for their meat and scales, mainly to meet the demands of **China** and **Vietnam**.
- Two poachers were caught while trying to sell an Indian Pangolin in **Vashi** on Wednesday, police said.

- While the price of 1 kilogram (2.2 pounds) of pangolin scale was below **$100** in 2008, it had risen to about **$500** by January 2016.
- Once the scales reach China, they cost around **$2,500 a kg (or Rs 1.6 lakh)**.

- On Thursday, Vietnamese officials released 16 Sunda Pangolins that had been confiscated by **Forest Protection Department** officers in 2015 to a “safe, undisclosed location in Vietnam,” according to Vietnam news.
- Last year, **“People for Pangolin”, a U.K.-based NGO**, launched a social media campaign called #pumpit4pangolins.

- In 2017, local authorities seized their largest stockpile to date of **130 kilograms (287 pounds)** of pangolin scales, in Kalpitiya,
- Chinese customs officials have seized over **23 metric tons (23,000 kilograms or 50,000 pounds)** of pangolin scales after a year-long investigation, local Chinese media reported last week.

FIGURE 9 Example of four types of Named Entities extracted along with the sentence they appear in for the keyword ‘*Pangolin*’

species mentioned in the respective articles; MONEY: 8,659 (486); PERSON: 123,324 (569); ORG: 145,705 (569); QUANTITY: 15,161 (516); GPE: 113,561 (568) and CARDINAL: 97,669 (569).

4 | DISCUSSION

This study demonstrates the potential of using natural language processing and machine learning to identify relevant articles focusing on species of conservation concern and extract relevant information that can be used for further analyses. The proposed methods use a text-vectorising approach for deduplicating articles and successfully implement automatic labelling of training data for use in classification task to remove irrelevant articles. Importantly, the proposed methods allow for accurate and fast classification of online content. While in this study we focus on CITES Appendix I listed species and mine data on these species from online news and Twitter, the proposed methods allow for automated content analysis from multiple digital platforms at the same time and for inclusion of a larger number of species globally. We foresee the potential use of these methods by conservation scientists and practitioners interested in investigating human–nature interactions from digital text but also to extract information on the ecology of species and ecosystems from any digital corpora (Jarić et al., 2020).

As digital data are generated in a continuous manner across various platforms and due to its large volume and regularity, manual processing of the same data is cumbersome, time-consuming and expensive (Stonebraker et al., 2013). This necessitates an automatic method to not only collect such data but also to automatically check for quality control. For this purpose, the primary tool we developed is a neural network to analyse the content of the articles and classify them as relevant or irrelevant. With a high classification performance during training/testing on the automatically annotated news data and a qualitatively good performance on a different domain data of the Google News articles, we demonstrated the effectiveness of this method on both counts of the efficacy of implementing a neural network and the annotation method for this particular task. Along with this, we implemented a vectorised deduplication method and language checking to further enhance the quality of the data we collect. Preliminary results indicate the success of this method, as it allowed collecting over 15,000 articles focusing on 585 different species along with some specific trade-related articles from Twitter sources (as indicated by topic modelling). There is a heavy bias for animal species compared to plants and within animals a bias towards mammals followed by birds and reptiles. We suspect this to be a combined effect of both the representation of species in the CITES list and the occurrence statistics on the web. Article ratios in Table 4 clarify the disparity of article occurrence across different classes of animal species. The popularity of charismatic species is reflected in the high number of articles dedicated to them, although some globally lesser known species are also included. Overall, this indicates a need to create more awareness about other species and reduce biases in popular media for certain species. Interestingly, the focus on large mammals also reflects a common bias in conservation science (Di Marco et al., 2017).

The pipeline we developed is highly modular and flexible and each stage can be adapted and modified for specific research questions. The species list can be extended to include all species listed in all CITES Appendices (i.e. also to Appendices II and III) as well as to, for example, all species listed in the IUCN Red List. The same methods can be applied to many more digital platforms (e.g. e-commerce platforms, many social media platforms, etc.) and also to text content in peer-reviewed articles and reports. As explained above, the modular nature of the pipeline allows building a small module to access data from a digital platform using their API or web-scraping (see e.g. Singrodia et al., 2019 for an overview of general web-scraping tools) and simply 'plug it' into the pipeline as an additional source. The output of the new module needs to be a list of hyperlinks to the articles extracted from the source. The resulting databases that include information collected over a long time period can be then used to explore multiple research questions in conservation science. The data collected through the pipeline can be used for a number of analyses, including, for example, (a) topic modelling (similar to Section 3.2) and sentiment analysis (Fink et al., 2020) of the textual content; (b) spatiotemporal analysis of wildlife trade volumes (each article is tagged with a time of publishing and may have several temporal markers in text); (c) spatial analyses of wildlife trade routes (to detect emerging or existing routes similar to the work in Indraswari et al. (2020); (d) quantitative analyses of the prices and quantities involved in wildlife trade from the content extracted using NER; (e) social network analysis to investigate networks of stakeholders involved with wildlife trade and (f) enhancing training datasets, for example the generated database can be used to annotate the text for specific entities or categories and thus help build more specialised or refined machine learning models. The above-mentioned list is far from exhaustive and focuses particularly on wildlife trade, but is extendable depending on the research objectives under which the pipeline will be deployed. Analysing online news content can also help increase awareness about the global biodiversity crisis in the public (Hooykaas et al., 2020).

As the filtering process is automated, there are errors that can potentially create noise in the data collected. We have attempted to reduce this noise as much as possible and as we collect more data and train our models further, we expect to improve upon the signal-to-noise ratio. Noise can have several origins, including, for example, (a) false positives from the model resulting in some articles not being conservation related; (b) the NER model making errors in detecting the type of entity or missing some entities and (c) spurious text present in the articles depending on website design that can create errors in the scraper or vectorisation techniques. A mitigating solution to such noise lies largely in the ability to have refined annotation of the training data. Another caveat to keep in mind is that the deduplication process effectively removes all identical articles, but not articles discussing the same topic, but using different vocabularies and/or differing in some additional details in text. Vectors arising from these articles will naturally be quite different from articles that are syndicated copies; thus, original articles will not be removed. However, we believe this information may be useful for additional analysis of the coverage, for example, in gauging differences in opinions, sentiments or facts about same topics. Although the model uses a neural network and statistical models, there is no strict requirement of special computational equipment like Graphical Processing Units (GPU) and standard CPUs can be used, albeit with some loss of speed, thus making the method applicable in wider scenarios. An important next step will be to combine automated text and image analyses together. As it stands, the pipeline extracts only news articles publicly available and intended for public dissemination and not information of intrinsically private nature such as closed groups. Still, accessing such closed groups without permission raises ethical concerns that may require minimum considering data privacy requirements (Di Minin et al., 2021; Zimmer, 2010).

In conclusion, our proposed pipeline may facilitate the extraction of digital data for further analyses, potentially increasing the efficiency at which this information can be extracted by conservation scientists and practitioners who may still be retrieving this information manually. These methods could be used to speed up data collection as part of efforts of creating online databases for species of conservation concern. As such, the pipeline serves as a first step in reducing the manual workload of researchers in collecting large amount of data. Future work should seek further refinement of the models by enhancing the annotation scheme and improving the filtering so that only information related to certain topics (e.g. wildlife trade) is retrieved and should also combine textual and image content analysis. With an aim to develop collaborative efforts and aid research in this domain, we provide the original code and distribute the data upon request for scientific exploration purposes (see Data Availability Statement).

ACKNOWLEDGEMENTS

R.K. and E.D.M. thank the European Research Council (ERC) for funding under the European Union's Horizon 2020 research and innovation programme (grant agreement 802933).

AUTHORS' CONTRIBUTIONS

R.K. and E.D.M. designed the research; R.K. developed and programmed the natural language processing and machine learning models; R.K. and E.D.M. wrote the paper.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13608.

DATA AVAILABILITY STATEMENT

Database and the source code are available at: Etsin (https://doi.org/10.23729/7e5c881b-0c80-4edf-b88a-0051b2f63ca0).

ORCID

Ritwik Kulkarni https://orcid.org/0000-0002-1320-9693
Enrico Di Minin https://orcid.org/0000-0002-5562-318X

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (pp. 265–283). USENIX Association. Retrieved from https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

Banker, K. (2011). *MongoDB in action*. Manning Publications Co.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media Inc.

Bombieri, G., Nanni, V., Delgado, M. D. M., Fedriani, J. M., López-Bao, J. V., Pedrini, P., & Penteriani, V. (2018). Content analysis of media reports on predator attacks on humans: Toward an understanding of human risk perception and predator acceptance. *BioScience*, *68*(8), 577–584.

Butchart, S. H., Walpole, M., Collen, B., Van Strien, A., Scharlemann, J. P., Almond, R. E., Baillie, J. E., Bomhard, B., Brown, C., Bruno, J., Carpenter, K. E., Carr, G. M., Chanson, J., Chenery, A. M., Csirke, J., Davidson, N. C., Dentener, F., Foster, M., Galli, A., … Watson, R. (2010). Global biodiversity: Indicators of recent declines. *Science*, *328*(5982), 1164–1168.

Corbett, J. B. (2016). When wildlife make the news: An analysis of rural and urban north-central us newspapers. *Public Understanding of Science*, *4*(4), 397–410.

Correia, R. A., Ladle, R., Jarić, I., Malhado, A. C. M., Mittermeier, J. C., Roll, U., Soriano-Redondo, A., Veríssimo, D., Fink, C., Hausmann, A., Guedes-Santos, J., Vardi, R., & Di Minin, E. (2021). Digital data sources and methods for conservation culturomics. *Conservation Biology*, *35*(2), 398–411. https://doi.org/10.1111/cobi.13706

Dahlgaard, S., Knudsen, M., & Thorup, M. (2017). Practical hash functions for similarity estimation and dimensionality reduction. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (pp. 6615–6625). Curran Associates Inc. Retrieved from https://dl.acm.org/doi/abs/10.5555/3295222.3295407

Di Marco, M., Chapman, S., Althor, G., Kearney, S., Besancon, C., Butt, N., Maina, J. M., Possingham, H. P., Rogalla von Bieberstein, K., Venter, O., & Watson, J. E. (2017). Changing trends and persisting biases in three decades of conservation science. *Global Ecology and Conservation*, *10*, 32–42.

Di Minin, E., Fink, C., Hausmann, A., Kremen, J., & Kulkarni, R. (2021). How to address data privacy concerns when using social media data in conservation science. *Conservation Biology*, *35*(2), 437–446.

Di Minin, E., Fink, C., Hiippala, T., & Tenkanen, H. (2019). A framework for investigating illegal wildlife trade on social media with machine learning. *Conservation Biology*, *33*(1), 210–213.

Di Minin, E., Fink, C., Tenkanen, H., & Hiippala, T. (2018). Machine learning for tracking illegal wildlife trade on social media. *Nature Ecology & Evolution*, *2*(3), 406–407.

Di Minin, E., Tenkanen, H., & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, *3*, 63. https://doi.org/10.3389/fenvs.2015.00063

Fink, C., Hausmann, A., & Di Minin, E. (2020). Online sentiment towards iconic species. *Biological Conservation*, *241*, 108289. https://doi.org/10.1016/j.biocon.2019.108289

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In D. E. Losada & J. M. Fernández-Luna (Eds.), *European conference on information retrieval* (pp. 345–359). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-540-31865-1_25#citeas

Han, J., Haihong, E., Le, G., & Du, J. (2011). Survey on nosql database. In *2011 6th international conference on pervasive computing and applications* (pp. 363–366). IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/6106531

Hernandez-Castro, J., & Roberts, D. L. (2015). Automatic detection of potentially illegal online sales of elephant ivory via data mining. *PeerJ Computer Science*, *1*, e10. https://doi.org/10.7717/peerj-cs.10

Hooykaas, M. J., Schilthuizen, M., & Smeets, I. (2020). Expanding the role of biodiversity in laypeople's lives: The view of communicators. *Sustainability*, *12*(7), 2768. https://doi.org/10.3390/su12072768

Indraswari, K., Friedman, R. S., Noske, R., Shepherd, C. R., Biggs, D., Susilawati, C., & Wilson, C. (2020). It's in the news: Characterising indonesia's wild bird trade network from media-reported seizure incidents. *Biological Conservation*, *243*, 108431. https://doi.org/10.1016/j.biocon.2020.108431

Jarić, I., Correia, R. A., Brook, B. W., Buettel, J. C., Courchamp, F., Di Minin, E., Firth, J. A., Gaston, K. J., Jepson, P., Kalinkat, G., Ladle, R., Soriano-Redondo, A., Souza, A. T., & Roll, U. (2020). iEcology: Harnessing large online resources to generate ecological insights. *Trends in Ecology & Evolution*, *35*(7), 630–639. https://doi.org/10.1016/j.tree.2020.03.003

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Kulkarni, R., Vintró, M., Kapetanakis, S., & Sama, M. (2018). Performance comparison of popular text vectorising models on multi-class email classification. In K. Arai, R. Bhatia, & S. Kapoor (Eds.), *Proceedings of SAI intelligent systems conference* (pp. 567–578). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-01054-6_41

Ladle, R. J., Correia, R. A., Do, Y., Joo, G.-J., Malhado, A. C., Proulx, R., Roberge, J.-M., & Jepson, P. (2016). Conservation culturomics. *Frontiers in Ecology and the Environment*, *14*(5), 269–275. https://doi.org/10.1002/fee.1260

Lamba, A., Cassey, P., Segaran, R. R., & Koh, L. P. (2019). Deep learning for environmental conservation. *Current Biology*, *29*(19), R977–R982. https://doi.org/10.1016/j.cub.2019.08.016

Marshall, B. M., Strine, C., & Hughes, A. C. (2020). Thousands of reptile species threatened by under-regulated global trade. *Nature Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-18523-4

Maxwell, S. L., Fuller, R. A., Brooks, T. M., & Watson, J. E. (2016). Biodiversity: The ravages of guns, nets and bulldozers. *Nature News*, *536*(7615), 143. https://doi.org/10.1038/536143a

Morcatty, T. Q., Bausch Macedo, J. C., Nekaris, K.-A.-I., Ni, Q., Durigan, C. C., Svensson, M. S., & Nijman, V. (2020). Illegal trade in wild cats and its link to chinese-led development in central and south america. *Conservation Biology*, *34*(6), 1525–1535.

Nghiem, L. T., Papworth, S. K., Lim, F. K., & Carrasco, L. R. (2016). Analysis of the capacity of google trends to measure interest in conservation topics and the role of online news. *PLoS ONE*, *11*(3), e0152802. https://doi.org/10.1371/journal.pone.0152802

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(25), E5716–E5725. https://doi.org/10.1073/pnas.1719367115

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, M., & Vrgoč, D. (2016). Foundations of JSON schema. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 263–273). International World Wide Web Conferences Steering Committee. Retrieved from https://dl.acm.org/doi/10.1145/2872427.2883029

Pimm, S. L., Alibhai, S., Bergl, R., Dehgan, A., Giri, C., Jewell, Z., Joppa, L., Kays, R., & Loarie, S. (2015). Emerging technologies to conserve biodiversity. *Trends in Ecology & Evolution*, *30*(11), 685–696. https://doi.org/10.1016/j.tree.2015.08.008

Ramos, D., Franco-Pedroso, J., Lozano-Diez, A., & Gonzalez-Rodriguez, J. (2018). Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy*, *20*(3), 208.

Régin, J.-C., Rezgui, M., & Malapert, A. (2013). Embarrassingly parallel search. In C. Schulte (Ed.), *International conference on principles and practice of constraint programming* (pp. 596–610). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-40627-0_45

Richardson, L. (2007). *Beautiful soup documentation*. April.

Roesslein, J. (2020). *Tweepy: Twitter for python!* Retrieved from https://github.com/tweepy/tweepy

Salton, G. (1991). Developments in automatic text retrieval. *Science*, *253*(5023), 974–980.

Singrodia, V., Mitra, A., & Paul, S. (2019). A review on web scrapping and its applications. In *2019 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1–6). IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/8821809

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Stonebraker, M., Bruckner, D., Ilyas, I. F., Beskales, G., Cherniack, M., Zdonik, S. B., Pagan, A., & Xu, S. (2013). Data curation at scale: The data tamer system. In *6th biennial Conference on Innovative Data Systems Research (CIDR)* (Vol. 2013). Citeseer. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.8817&rep=rep1&type=pdf

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, *233*, 298–315. https://doi.org/10.1016/j.biocon.2019.01.023

Toleu, A., Tolegen, G., & Mussabayev, R. (2020). Deep learning for multilingual pos tagging. In M. Hernes, K. Wojtkiewicz, & E. Szczerbicki (Eds.), *International conference on computational collective intelligence* (pp. 15–24). Springer. https://doi.org/10.1007/978-3-030-63119-2_2

Unger, S. D., & Hickman, C. R. (2020). A content analysis from 153 years of print and online media shows positive perceptions of the hellbender salamander follow the conservation biology. *Biological Conservation*, *246*, 108564.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Zimmer, M. (2010). 'but the data is already public': On the ethics of research in facebook. *Ethics and Information Technology*, *12*(4), 313–325. https://doi.org/10.1007/s10676-010-9227-5

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Kulkarni R, Di Minin E. Automated retrieval of information on threatened species from online sources using machine learning. *Methods Ecol Evol*. 2021;12:1226–1239. https://doi.org/10.1111/2041-210X.13608